Intelligence 56 (2016) 1–9

Contents lists available at ScienceDirect

Intelligence

# Speed and accuracy indicators of test performance under different instructional conditions: Intelligence correlates

Phillip L. Ackerman *, Victor J. Ellingsen

*Georgia Institute of Technology, USA*

ARTICLE INFO

*Article history:*
Received 13 December 2015
Received in revised form 31 January 2016
Accepted 12 February 2016
Available online xxxx

*Keywords:*
Intelligence
Speed
Accuracy
Instructions

ABSTRACT

Performance on speeded ability tests, in contrast to power tests, reflects an individual's ability to get answers correct and to do so rapidly. However, with speeded tests, overall performance scores represent some unknown combination of the individual's strategy toward greater speed or higher accuracy. Scoring methods, such as penalties for wrong answers, are often imposed to either encourage examinees to adopt a specific speed-accuracy tradeoff strategy, or to attempt to derive performance scores that 'factor out' such strategic differences. In the current study, baseline assessments of four perceptual speed and psychomotor ability tests were administered, along with three different instructional conditions (accuracy-emphasis, speed-emphasis, and balanced accuracy and speed). A general ability composite was derived from a battery of intellectual ability tests. Changes in speed-accuracy tradeoff emphasis resulted in a consistent pattern of changes in the *g* correlates of latency/completion time performance indicators and number of errors. Increasing emphasis on accuracy resulted in increasing *g* correlates with latency/completion time, and decreases in *g* correlates with error rates. Implications for construct validity of ability tests and for further consideration of the conditions of testing are presented.



## 1. Background

From the early years of modern testing, there has been a robust debate about the importance of *speed* and *level* of performance on ability tests, as they relate to the overall construct of intelligence. Thorndike, Bregman, Cobb, and Woodyard (1926), for example, noted that "*Other things being equal, the more quickly a person produces the correct response, the greater is his* [sic] *intelligence*" (p. 24, italics in original). That is, Thorndike considered speed to represent an intellectual advantage when individuals were equivalent on "level." In practice, however, it is difficult or impossible to have examinees perform at an equal level, but differ only on speed.

Early in testing research and applications, the main concern of investigators was whether or not increasing testing time to reduce the speed-demands of tests resulted in differential effects on the rank-ordering of individuals on test performance (e.g., Beck, 1933; Tinker, 1931 for early reviews). The results of such investigations were generally mixed—some tests with different speed constraints yielded correlations close to the individual test reliabilities, while others did not. Thorndike, in fact, recognized that individual differences in time management could result in higher or lower test scores, but he concluded that such differences were not of substantial importance (Thorndike et al., 1926).

Later investigators were better aware of the potential measurement issues associated with the fact that individuals could *trade-off* accuracy for speed on some tests. That is, there is a fundamental dependence between how rapidly one completes test items and how accurate (correct) the answers are. Most researchers attempted to grapple with this issue by trying to equalize performance of individuals who adopted different speed-accuracy tradeoff strategies. For example, discussions have been accorded to determining the 'best' (from reliability and validity perspectives) methods for computing the total score for multiple-choice tests. Numerous alternate formulas have been proposed. Such formulas typically either provide no credit for incorrect responses or impose a penalty in an attempt to 'correct for guessing' on other items for which one can reasonably infer the examinee did not actually know the correct answer, but managed to guess correctly (e.g., see Lord, 1975, for a review; see also Diamond & Evans, 1973).

When test items vary considerably in terms of difficulty, however, there are multiple causes of wrong answers to test items. The examinee may be running out of time on a timed test with multiple items uncompleted, so the examinee may randomly guess the answers on the remaining test items. If that is the case, then a correction for guessing that takes account of such random responding might subtract a fraction of a point for each incorrect response (where the fraction is equal to one divided by the number of item response options). If the test developer wishes to discourage guessing entirely, then a subtraction for wrong answers that is larger than one over the number of item response options would presumably yield an appropriate penalty. Yet, in operational situations, examinees often have partial knowledge about the correctness

* Corresponding author at: School of Psychology, Georgia Institute of Technology, 654 Cherry St., MC 0170, Atlanta, GA 30332-0170, USA.
*E-mail address:* plackerman@gatech.edu (P.L. Ackerman).

http://dx.doi.org/10.1016/j.intell.2016.02.004
0160-2896/

or incorrectness of response options. In such cases, when penalties for wrong answers are known, the examinee might evaluate how many wrong options can be eliminated from consideration, and then determine whether the chances of getting the item wrong are lower than the penalty for guessing—leading to a decision of whether a partially-informed guess for that item is worth the risk.

For constructed response test items (e.g., sentence completion tests), however, the chances for successful guessing are generally unknown, but often are believed to be very low. For these kinds of tests, developers typically see little need to correct scores for guessing, and simply sum the number of correct answers to obtain a final test score. Similarly, for power tests (which are traditionally untimed), investigators have largely ignored errors, and focused only on the number of correct answers. However, it is not clear whether salient individual differences exist in a propensity to take short-cuts by guessing about responses, when the examinee does not recognize the correct answer.

For speeded tests, in which items are sometimes relatively homogeneous in difficulty, such as tests of basic math (subtraction and addition), perceptual speed, and psychomotor abilities—the assumption that examinee 'knowledge' will mainly determine item response accuracy is generally untenable. This is primarily because the expectation is that with unlimited time and a modicum of motivation (e.g., see Thurstone, 1937), test performance would be error-free. For such tests, the underlying perspective is that there is a speed-accuracy tradeoff function for each individual, depending on his/her ability level (e.g., see Lohman, 1989a). When intelligence investigators first adopted the information processing perspective for ability testing in particular, there was an active controversy about whether speed of responding to simple items is a salient indicator of intellectual ability (e.g., see Jensen, 1998, 2006; Vernon, Nador, & Kantor, 1985; though see Sternberg, 1986 for an opposing viewpoint).

The problem of speed-accuracy tradeoff (Wickelgren, 1977) is that the function that describes the relationship is generally found to be decidedly non-linear. Accuracy increases rapidly with decreasing speed initially, but as accuracy reaches asymptotic levels, latency tends toward infinity (i.e., diminishing returns on accuracy with increasing latency). Performance on most group ability tests is the total number of items correctly answered (or total number of items answered correctly minus some fraction of the total number of items marked incorrect). However, for speeded tests, the total number of items answered correctly is the result of an unknown combination of the examinee's knowledge and his or her ability or motivation to answer items quickly. Compared to lower ability examinees, higher ability examinees are presumed to respond more accurately at a given response speed, or with shorter response latencies at a given level of accuracy, in line with Thorndike et al.'s (1926) suggestion. However, the precise shape of each individual examinee's speed-accuracy tradeoff function is generally unknown. This makes it impossible to determine each individual's location on his or her own speed-accuracy tradeoff function, based only on a single latency and error score. Therefore, in most testing situations, it is generally impossible to determine which of two individuals has higher ability, unless one individual is *both* faster *and* more accurate than the other individual. Furthermore, because the shape of the speed-accuracy tradeoff function is unknown and likely nonlinear under normal testing conditions, it is not feasible to derive an accurate 'correction for guessing' for speeded tests that is as straightforward as that used for traditional multiple-choice tests. Some approaches to determining speed-accuracy tradeoff functions have been tried, such as imposing different response deadlines or varying the amount of time an item is displayed, most notable is the work by Lohman (1986, 1989a, 1989b) on spatial visualization tests. Other approaches have examined the reliability/consistency of tests, and the reactivity of individuals, when confronted with tests that are administered with explicit manipulation of payoffs for right and wrong answers (e.g., Quereshi, 1960). Nonetheless, many tests retain a simple fractional penalty for guessing, even when the items are homogeneous in difficulty levels, such as tests of perceptual speed ability (e.g. see Ekstrom, French, Harman, & Dermen, 1976).

Numerous studies have attempted to address the implications of speed-accuracy tradeoffs for ability testing, whether in terms of: (a) providing explicit instructions regarding accuracy and errors; (b) deriving test scores that attempt to equate scores of individuals who appear to differ on an emphasis of speed or accuracy; (c) exploring derivations of individual speed-accuracy tradeoff functions; or (d) manipulating speed-accuracy tradeoff by changing payoff matrices for correct answers and errors. However, little is known about how changes in instructional emphasis for speed vs. accuracy affect the construct validity of tests, especially in terms of the *g*-loading of performance measures. This question is important because tests that were designed to be administered under relatively liberal time limits are sometimes administered under speeded conditions in research settings (such as Raven's Progressive Matrices in some working memory research; for a review and meta-analysis, see Ackerman, Beier, & Boyle, 2002; Ackerman, Beier, & Boyle, 2005). This question is difficult to address using power tests containing items of varying levels of difficulty, because speed/accuracy tradeoff will be confounded with item difficulty, making the results difficult to interpret. However, tests composed of items of uniform difficulty allow more straightforward manipulation of the speed-accuracy tradeoff and interpretation of results. The current study focused on a small set of such tests and a set of reference ability measures, in order to investigate changes in *g*-loadings in response to speed-accuracy instructional manipulations.

## 2. Current study

In the current study, the main question of interest was whether different instructional emphases on speed or accuracy influence the *g*-loading of various tests, and whether these instructional effects were reflected in separate latency and accuracy scores for the tests. To explore this question, we selected a set of perceptual speed and psychomotor tests, mainly because they have two advantages over other kinds of tests. One principal advantage of these tests is that they typically involve items of roughly equal difficulty, which means that the underlying speed/accuracy function for each item is essentially equivalent. The second advantage is that, in contrast to traditional fluid or crystallized intelligence tests, the test items can be answered correctly without error in untimed administrations (e.g., see Ackerman, 1988). Thus, errors on such tests are entirely a result of the examinees' internal rules or experimenter-imposed instructions for speed-accuracy emphases. However, one disadvantage to such tests that must be noted is that there are typically significantly larger 'practice' effects, compared to crystallized or fluid intellectual ability tests. The presence of practice effects means that one should either provide initial practice (to improve the stability of test performance) on the tests prior to instructional manipulations, or counterbalance the order of instructional manipulations (or both), to avoid carry-over confounds from order effects.

Tests that make demands on perceptual speed and psychomotor abilities—especially those that include more complex processing than simple reaction time or cancellation tests—have historically been included in omnibus intelligence tests (e.g., the digit-symbol subtest in the Wechsler Scales). Complex perceptual speed and psychomotor tests often correlate substantially with other measures of general intelligence, especially those that are speeded to some degree (e.g., see Ackerman & Cianciolo, 1999; Alderton, Wolfe, & Larson, 1997; Allison, 1960; Melton, 1947). These tests also often have significant incremental predictive validity (beyond non-speeded *g* measures) for occupational performance, such as in aviation and dentistry domains (Ackerman & Kanfer, 1993). Nonetheless, modern intelligence test batteries have removed these tests—not because of a lack of construct or criterion-related validity, but because pre-computer versions of such tests often required specialized apparatus for administration (Fleishman, 1953).